**HUYNH DO**

**Module 3, Part 1/2**

1. **Specify which variable is the dependent and independent variable**
* **AGE**: This is the **independent** variable because as we are getting older, the pancreas sometime has a difficult time keeping blood sugar levels in the optimal range; therefore, it is used to explain the cause of diabetes variables rather than being influenced by them.
* **DIABETES**: This is the **dependent** variable because it represents the outcome or result that is being studied in relation to other factors, in this case that is age.
1. **Using the variables, state the null and alternative hypotheses.**

With specified both variables, we can construct these 2 hypotheses:

1. **Null Hypothesis (H₀):**

There is **no relationship** between age and whether the participants report that they have any type of diabetes. In other words, aging is **not** a factor in diabetes disease.
**H₀**: Age and having diabetes are **not** associated.

1. **Alternative Hypothesis (H₁)**

There is a relationship between age and diabetes when get older.

**H₁**: Age and being told diabetes are dependent.

1. **Perform the Chi-Square test of independence and interpret the results.
**

****

****

* **Chi-Square Observation**

****

1. **The p-value**
* Pearson Chi-Square: The value is 144.861 with 65 degree freedom (df)
* **The p-value** (Asymptotic Significance) came up with < 0.001, which is much smaller than the standard 0.05 level.
1. **Likelihood Ratio:**
* The value is **134.001** with 65 degrees of freedom (df)
* Both p-value for Likelihood Ratio values are < .001 suggests a strong evidence against the null hypothesis (**H₀**)
1. **Linear-by-Linear Association:**
* The **Linear-by-Linear Association** value (51.568, p < 0.001) suggests that, as age increases, the likelihood of being suffered from the diabetes.
1. **Foot note**
* The foot note indicates that 78 cells or 59.1% have expected counts less than 5 combine with minimum expected is 0.11 suggest that some cells in the contingency table have very small frequencies which may violate assumption of the Chi-Square test.
* **Interpretation of Results and Findings**
* **Data Significance**:
* Since both Pearson Chi-Square test and Likelihood Ratio have p-value significantly < 0.001, it is strong suggested that we should **reject** the null hypothesis (**H₀**). Therefore, we can safely conclude that there is an association between being age and being told diabetes.
* **What is the strength of association**:
* Both high Chi-Square values (**144.861** and **134.011**) indicate that the relationship between the 2 variables is very strong, meaning that the chance of having diabetes increase when aging is quite high.
* **Findings:**
* There is a significant association between the age of respondents and whether they have been told they have diabetes. However, a large portion of the cells have expected less than 5 combined with a large missing case (40.1%) might compromise the level of confidence, therefore, further investigation might be needed to boost the validity of this finding.
1. **Assessing the Assumptions of the Chi-Square Test:**

To ensure the highest confidence possible, the Chi-Square test must satisfy these key assumptions:

1. **Independence of Observations:**
* **Assumption: Individual participant should be truthful about the age and current condition, and more importantly, each responder is completely independent of others. In another word,** one person’s answer about age or diabetes should not manipulate or being influenced by another person’s answer.
* **Assessment:** In this case, each participant provides an individual response about their age and diabetes status, so we can assume independence, therefore, that data meets this assumption.
1. **Sufficiently Large Sample Size:**
* **Assumption: The sample size is expected to be in a large quantity to make sure the Chi-Square is valid.**
* **Assessment:** In this case, although, the total valid cases are 1,407 or 59.9% out of 2,348 which is an acceptable sample size, however almost half of the cases (40.1%) are missing raises the concern about the level of confidence of the outcomes.
1. **Expected Frequency in Each Cell:**
* **Assumption: The expected frequency in each cell in the contingency table should have a minimum 5 counts. If any cells that does not meet the requirement, the Chi-Square test may not valid.**
* **Assessment:** In this case, from the foot note, there are 78 cells (59.10%) have expected counts less than 5, and the minimum expected count is **0.11,** which is a big violation of the Chi-Square assumption.
1. **If necessary, perform post-hoc tests and interpret the results.**

The statistical data and findings indicate a significant association between age and being told diabetes. With the extreme Chi-Square result (p < 0.001), it is necessary to conduct this test to understand further how strong the association is.

The Standardized Residual test will be used for assessing the adequacy of statistical models and guiding model refinement.

To come to a conclusion, let’s calculate Standardized Residual at couple data points bases on group of ages such as Young Adults (23-31), Middle-Aged (40-50), and Senior (86+)

1. The Standardized Residual formula:
Expect Counts **(E) = (**Row Total × Column Total for ’Diabetes: Yes’)**/**Grand Total



* Where O is the observed count for "Diabetes: Yes"
* E is the expected count
1. With the formula above and from the contingency table, the Standardized Residual of some key turning ages are calculated as follow:
* Age 23: E = 2.5 | R = -1.58
* Age 31: E = 2.72 | R = -1.04
* Age 40: E = 3.26 | R = -0.70
* Age 50: E = 3.26 | R = -0.14
* Age 60: E = 2.61 | R = 1.48
* Age 86: E = 0.109 | R = 2.70
* Age 89 or Older: E = 0.109 | R = 2.70

From the data grid above, let’s construct a Line Chart to visualize the trend of data to further understanding of how strong of the association between age and diabetes.



* **Interpretation**:
1. Expected Counts (E – blue line):
* The expected counts are relatively stable for 2 groups Young Adults (23-31), Middle-Aged (40-50), and the values fall with the range of [2.5 - 3.26]. This indicates a similar expected number of diabetes cases across these age groups.
* The senior group, the expected counts drastically drop close to 0, which is 0.109, reflecting the smaller sample size for this group.
1. Standardized Residuals (R – green line):
* Both Young Adults (23-31) and Middle-Aged (40-50) have negative residuals which indicating that fewer observed diabetes cases than expected.
* **Senior (86+):** The residual value rises significantly above 0 (**2.70**), this indicates that the observed diabetes cases among the elderly are much higher than expected.
* **Findings:**
1. **Increasing Risk When Aging:** The line chart shows the upward **R** pattern confirms that being told diabetes increases along the ages. Young adults seem to less worry about this since fewer cases than expected, while older adults exhibit a much higher prevalence.
2. **Sharp Increase in Older Age Groups:** The spike in R values, especially from age of 86 and above, suggests that diabetes becomes particularly common among the elderly, exceeding the expected rates significantly.
3. **Using a clustered bar chart, visually display the results of the Chi-Square test of independence and explain your findings.**



* **Interpretation**:
* **Young Adult (18-30):** This group has the most dominant counts with diagnosed with no diabetes, and most having zero diabetes-reported cases. So, in general, young adults usually don’t suffer from diabetes.
* **Middle Age Groups (31-50):** From age 31 to 50, the number of people who has diabetes starts to increase a little bit but not much. This number could be explain due to changes in lifestyle, weight gain etc..
* **Older Age Groups (50 and above):**Starting from the age of 50, the body starts to show signs of trouble, and the number of diabetes diagnoses as age rises fast and steadily, particularly from around age 54 and older.
* **Senior (80 and above):** Although the number of participants is quite low, however, there are still a few elderly who reported diabetes. The decrease in the number of cases could be due to their current health which would not allow them to participate in the survey.
* **Conclusions**

Overall, the data suggests a clear association between age and the likelihood of being told one has diabetes, with older age groups will expose to higher chance of having diabetes. This pattern supports the understanding that age is a significant risk factor for diabetes.

1. **Choose a third variable to use as a control variable and rerun the Chi-Square test of independence along with post-hoc tests. (PHYSACTS)**

To add more to the complexity, we will add an additional “To what extent R can accomplish everyday physical activities - PHYSACTS” control variable to see whether this variable has contribution to the relationship between age and "Told have diabetes" variables





1. **Results Interpretation:**
* **Completely**:
* The Chi-Square Test value is 100.01 with 61 degrees of freedom. Both Linear-by-Linear and p-values are very close to 0 (< 0.001) strongly suggesting a statistically significant association between diabetes status and the ability to perform everyday activities completely.
* **Mostly**
* The Chi-Square Test value = 61.27 with 57 degrees of freedom and a p-value = 0.325, indicating no significant association in this group. However, the Linear-by-Linear Association shows a significant relationship (p < 0.001), therefore, together, data suggests there could be effects to the association but not guarantee.
* **Moderately**:
* The Chi-Square Test value = 48.22 with 44 degrees of freedom and the p-value = 0.306 (a little close to 0) and Linear-by-Linear = 0.017, indicating somewhat significant association.
* **A little**:
	+ - The Chi-Square Test value = 15.60 with 19 degrees of freedom and a p-value = 0.68 combines with Linear-by-Linear = 0.52, indicating no significant association between diabetes status and activities "a little."
* **Not at all**:
* No Chi-Square Test performed for this category, it is likely because there’s only one response was recorded.
* **Total**:
* When combining all categories, the total Chi-Square Test value is 144.81 with 65 degrees of freedom, and the p-value is <0.001, indicating a significant association across all levels.
1. **Summary:**
* There is a statistically significant association between the ability to accomplish everyday physical activities and diabetes status when looking at the data as a whole (p < 0.001). In additional, statically, the association is clear in some categories ("Completely" and "Total") than in others.